Animaltracker Data Validation: New Mexico Data

Joe Champion, Boise State University, joechampion@boisestate.edu
Thea Sukianto, Boise State University

May 27, 2020

This document analyzes the results of the animaltracker package's data cleaning procedures by comparing a sample of 8 data sets processed by the R package to the same data manually processed with Microsoft Excel.

The cleaning process uses flag-based rules for discarding cases (rows) of data.

- If measured rate of travel exceeds 84 m/min, mark the case with a RateFlag.
- If course change exceeds 100 degrees, mark the case with a CourseFlag.
- If measured distance traveled exceeds 840 m, mark the case with a DistanceFlag.
- Only keep cases without a DistanceFlag AND less than 2 flags.

Note: Throughout this report, the suffix x indicates data cleaned by animaltracker, and y indicates manually cleaned data.

Preliminaries

For reproducibility, configure and load required R packages, including animaltracker.

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(animaltracker)
library(psych)
```

Read and Prepare Data

Load the **manually** cleaned data (reshaping for consistent column names), then directly read and process the raw data using the animaltracker app.

```
###  load MANUALLY cleaned data, reshape for consistency
clean_manual <- read.csv("nm_validate/MastersheetNM - combined corrections applied.csv",
                             stringsAsFactors = FALSE,
                             na.strings =c("", "#VALUE!", "NA", "#N/A","#DIV/0!" )) %>%
  filter(!is.na(Date), !is.na(Cow)) %>%
  rename(CourseDiff = coursedifference, TimeDiff = timedifference,
     TimeDiffMins = timedifference.in.minutes,
     RateFlag = ratestatement, CourseFlag = coursestatement,
     DistFlag = distancestatement, TotalFlags = total, Keep = statement) %>%
  mutate(
     Index = as.numeric(Index),
     Altitude = as.numeric(Altitude),
     DateTime = paste(Date, Time),
     Keep = 1*!Keep, # ew
     ## fix undefined / missing flags
     RateFlag = replace_na(RateFlag, 1),
     CourseFlag = replace_na(CourseFlag, 1),
     DistFlag = replace_na(DistFlag, 1),
     TotalFlags = ifelse(is.na(TotalFlags), RateFlag+CourseFlag+DistFlag, TotalFlags),
     Keep = replace_na(Keep, 0)
 )

### read and CLEAN the raw data with the animaltracker app
folder_rawdata <- "../test_data/DeepWell_2018_Collar_Raw"
nm_files <- list.files(folder_rawdata)
aniid <- as.integer(gsub("DW_(\\d{3})(.*)", "\\1", nm_files))
gpsid <- as.integer(gsub("DW_(\\d{3})_(\\d{2})(.*)", "\\2", nm_files))

clean_anitracker <- lapply(1:length(nm_files), function(i){
    df_raw <- read.csv(file.path(folder_rawdata, nm_files[i]))
    df_clean_animaltracker <- clean_location_data(df_raw,
                                  dtype = "igotu", filters = FALSE, maxtime =150,
                                  aniid = aniid[i], gpsid = gpsid[i])
  }) %>%
  do.call(rbind, .) %>%
  rename(Cow = Animal) %>%
  type.convert()
```

Next, merge the cleaned data from `animaltracker` (167901 rows, 35 columns) with the manually cleaned data (167901 rows, 29 columns).

Rows are matched by the combination of `Cow`, `Index` (uniquely identifies almost all rows) and `Altitude` (to break ties in rare duplicates).

```
clean_anitracker <- clean_anitracker %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())

clean_manual <- clean_manual %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())

join <- full_join(clean_anitracker, clean_manual, by="merge_index") %>%
  rename( MegaRateFlag.x = MegaRateFlag) %>%
  mutate( Cow = factor(Cow.x))
```

The merged data has 167901 rows.

Analysis

Overall Agreement

First, we compare the results of cleaning the data within `animaltracker` (via the `clean_location_data` function) to results of manual cleaning via spreadsheet.

```
keepxtab <- with(join, table(Keep.x, Keep.y))
```

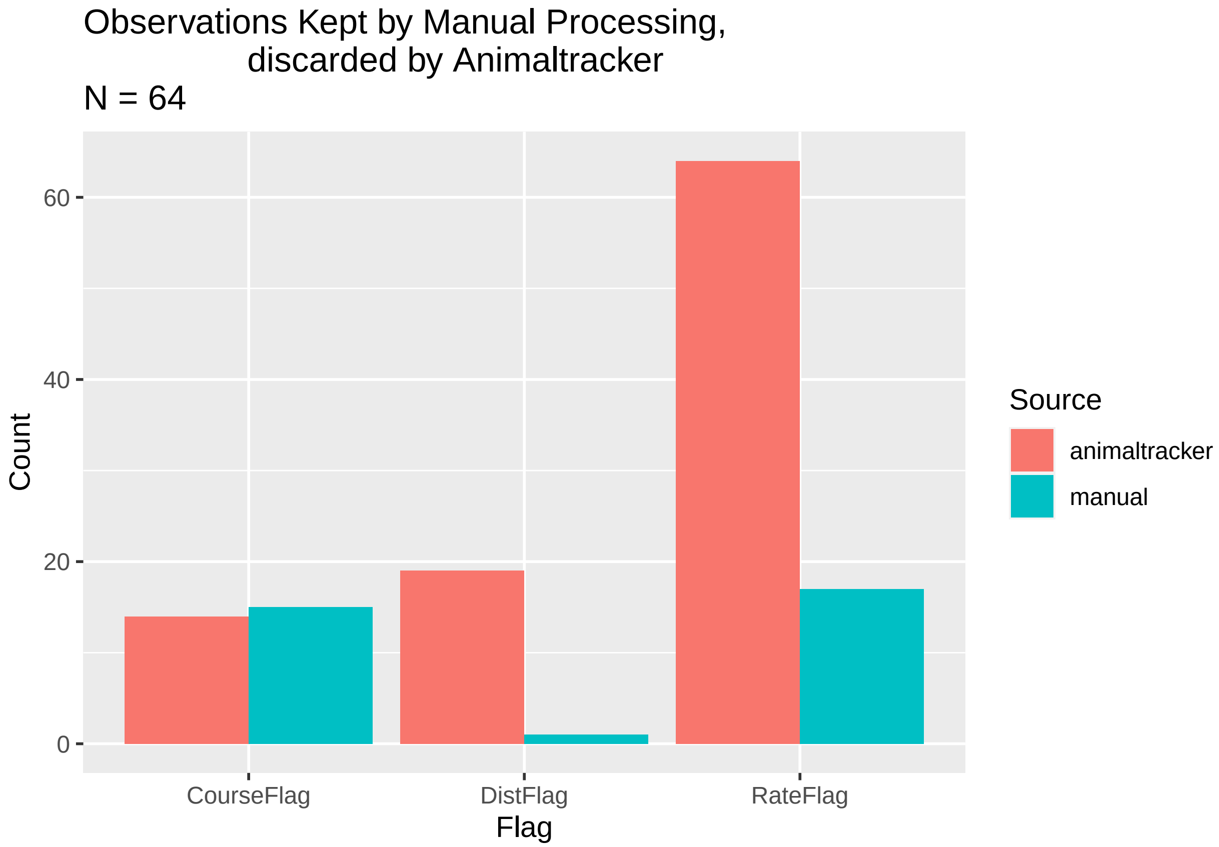
The cleaning methods agree in 99.958% of cases, except for 6 cases (0.004%) kept by `animaltracker` but discarded by manual processing and 64 cases (0.038%) kept by manual processing but discarded by `animaltracker`.

Analysis of Cases with Different Results

All of the cases kept by manual processing (n = 64) but discarded by `animaltracker` were marked with a **rate flag** by `animaltracker`, but not manual.

```r
manual_keep <- join %>%
  filter(Keep.x < Keep.y) %>%
  select(ind = merge_index, Cow, DateTime = DateTime.x, TimeDiffMins = TimeDiffMins.x,
              Rate.x, Rate.y, RateFlag.x, RateFlag.y,
              Dist.x = Distance.x, Dist.y = Distance.y, DistFlag.x, DistFlag.y,
              CourseDiff.x, CourseDiff.y, CourseFlag.x, CourseFlag.y)

manual_keep %>%
  summarise(RateFlag.x = sum(RateFlag.x),
                 CourseFlag.x = sum(CourseFlag.x),
                 DistFlag.x = sum(DistFlag.x),
                 RateFlag.y = sum(RateFlag.y),
                 CourseFlag.y = sum(CourseFlag.y),
                 DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
              Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle(paste0("Observations Kept by Manual Processing,
                discarded by Animaltracker\n","N = ",nrow(manual_keep)) )
```



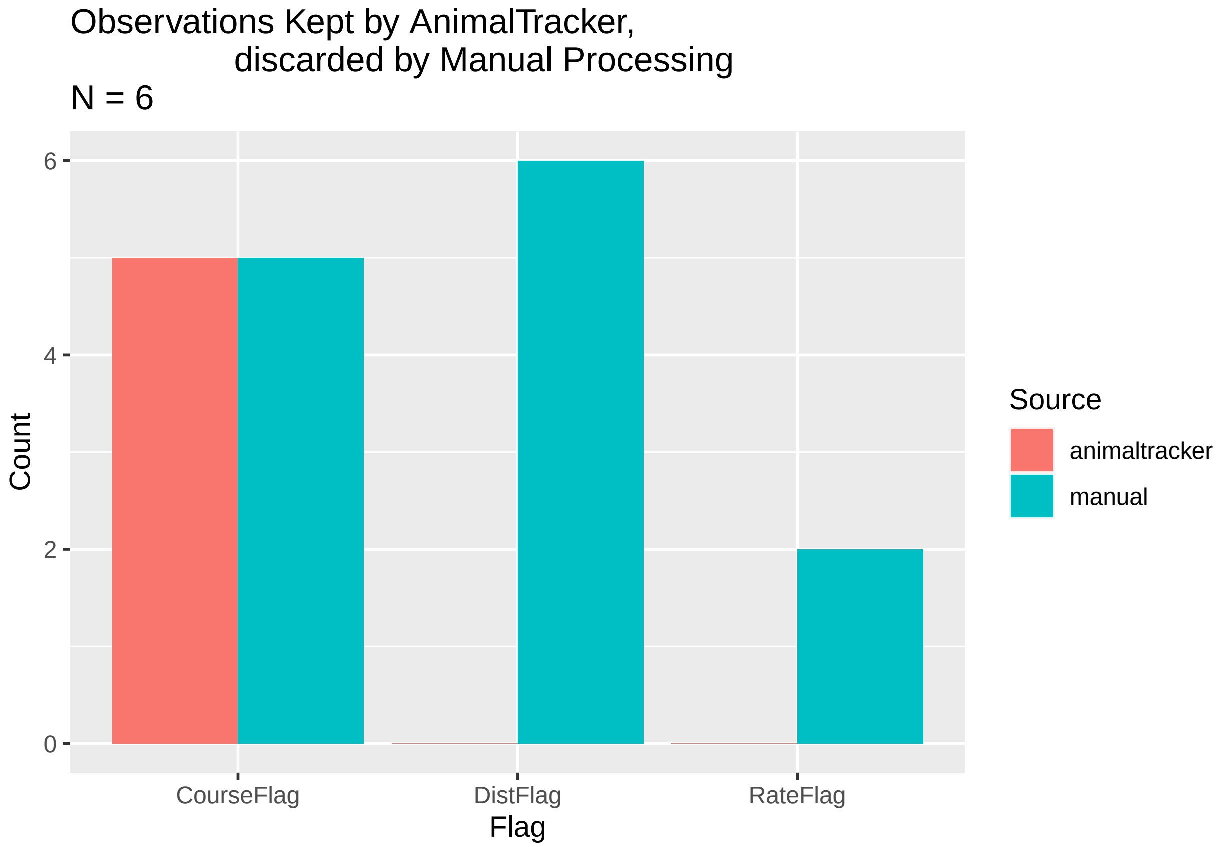
```
manual_keep %>% head(10) # first several cases
```

```
##      ind Cow            DateTime TimeDiffMins    Rate.x   Rate.y RateFlag.x
## 1      2  11 2018-05-23 15:48:22     0.100000 235.21787       NA          1
## 2  17071  11 2018-06-16 19:35:38     1.983333  84.52721 84.48521          1
## 3  17110  11 2018-06-16 20:53:10     1.983333  84.53694 84.37256          1
## 4  24006  63 2018-05-23 15:55:35     0.000000       NaN  0.00000          1
## 5  24029  63 2018-05-23 16:46:12     0.000000       NaN  0.00000          1
## 6  24030  63 2018-05-23 16:53:40     0.000000       NaN  0.00000          1
## 7  24031  63 2018-05-23 17:14:07     0.000000       NaN  0.00000          1
## 8  24090  63 2018-05-23 19:55:33     0.000000       NaN  0.00000          1
## 9  24091  63 2018-05-23 20:16:32     0.000000       NaN  0.00000          1
## 10 26022  63 2018-05-26 12:56:07     1.883333  86.17136 86.02308          1
##    RateFlag.y    Dist.x     Dist.y DistFlag.x DistFlag.y CourseDiff.x
## 1           0  23.48571  23.485714          1          0            0
## 2           1 167.56233 167.562333          1          0           23
## 3           1 167.33892 167.338916          1          0           28
## 4           0   0.00000          0          0          0            0
## 5           0   0.00000          0          0          0          183
## 6           0   0.00000          0          0          0            0
## 7           0   0.00000          0          0          0            0
## 8           0   0.00000          0          0          0            0
## 9           0   0.00000          0          0          0            0
## 10          1 162.01013 162.010131          1          0            6
##    CourseDiff.y CourseFlag.x CourseFlag.y
## 1             0            0            0
## 2            23            0            0
## 3            28            0            0
## 4             0            0            0
## 5           183            1            1
## 6             0            0            0
## 7             0            0            0
## 8             0            0            0
## 9             0            0            0
## 10            6            0            0
```

All of the cases kept by `animaltracker` but discarded by manual processing (n = 6) were marked with a **distance flag** by manual processing, but not `animaltracker`.

```
anitracker_keep <- join %>%
  filter(Keep.x > Keep.y) %>%
  select(ind = merge_index, Cow, DateTime = DateTime.x, TimeDiffMins = TimeDiffMins.x,
              Rate.x, Rate.y, RateFlag.x, RateFlag.y,
              Dist.x = Distance.x, Dist.y = Distance.y, DistFlag.x, DistFlag.y,
              CourseDiff.x, CourseDiff.y, CourseFlag.x, CourseFlag.y)

anitracker_keep %>%
  summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
              Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle(paste0("Observations Kept by AnimalTracker,
                 discarded by Manual Processing\n","N = ",nrow(anitracker_keep)) )
```



Observations Kept by AnimalTracker,
discarded by Manual Processing
N = 6

```
anitracker_keep %>% head(10) # first several cases
```

```
##      ind Cow            DateTime TimeDiffMins   Rate.x       Rate.y RateFlag.x
## 1   5024  11 2018-05-30 17:16:31     2.050000 83.93486     84.01382          0
## 2  39845  63 2018-06-14 21:32:10     2.100000 81.30984     81.47052          0
## 3  96488 257 2018-06-21 20:59:42     2.116667 82.15747     82.24029          0
## 4  99911 322 2018-05-23 23:02:07     0.100000  0.00000 907251.03920          0
## 5 119434 437 2018-05-23 18:41:44     2.133333 81.07113     81.15127          0
## 6 157563 535 2018-06-12 00:01:24     2.116667 82.23331     82.42689          0
##   RateFlag.y      Dist.x      Dist.y DistFlag.x DistFlag.y CourseDiff.x
## 1          1   172.2283  172.228339          0          1            3
## 2          0   171.0881  171.088102          0          1          112
## 3          0   174.0753  174.075285          0          1          314
## 4          1 90725.0966 90725.09661          0          1          236
## 5          0   173.1227  173.122707          0          1          160
## 6          0   174.4703  174.470257          0          1          353
##   CourseDiff.y CourseFlag.x CourseFlag.y
## 1            3            0            0
## 2          112            1            1
## 3          314            1            1
## 4          236            1            1
## 5          160            1            1
## 6          353            1            1
```

Effects of Cleaning Differences on Outcome Measures

The remaining analysis addresses the statistical effects of the processing errors on key outcomes.

Cumulative Distance (per day)

The time series plots below indicate a very close conformity between the data cleaned in `animaltracker` and the manually cleaned data.

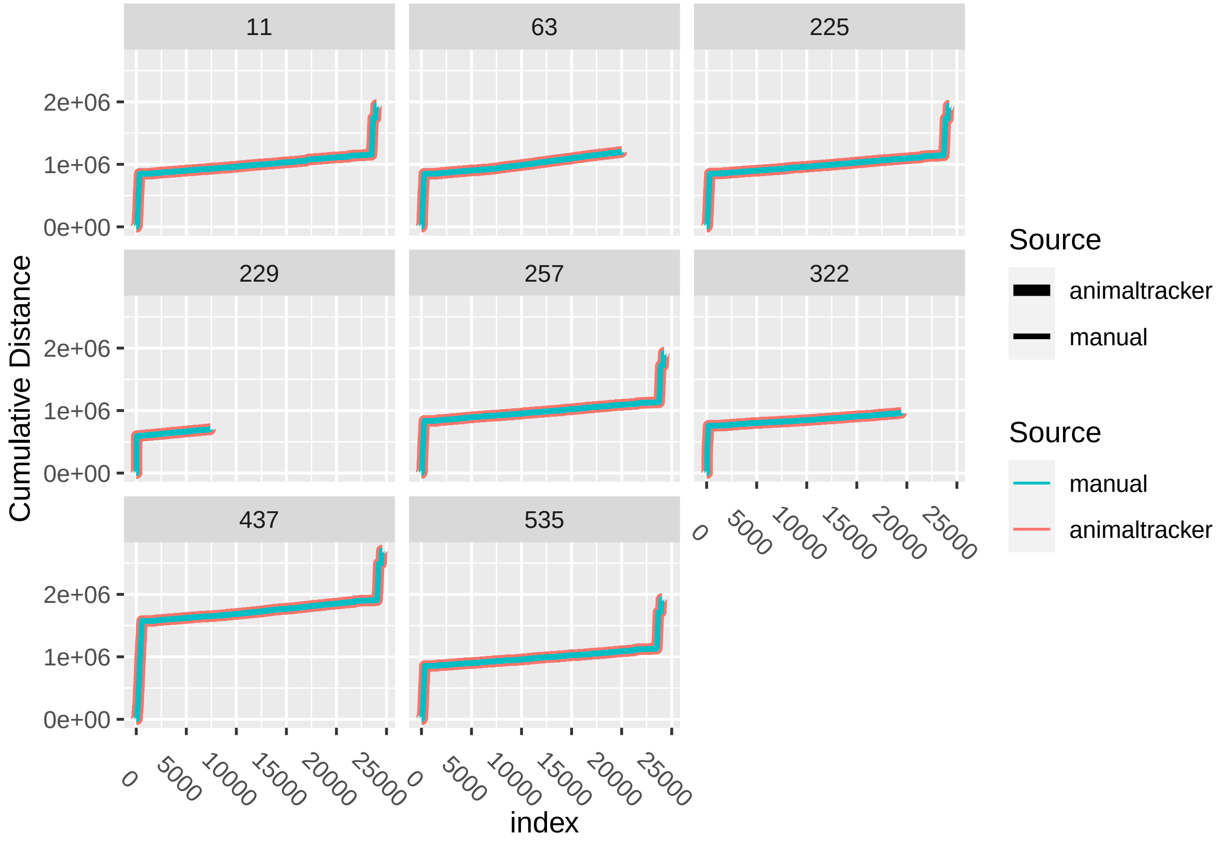
```
cumdist <- join %>%
  group_by(Cow) %>%
  arrange(merge_index) %>%
  mutate(Dist.y = lag(Distance.y,1),
                cumDist.x = cumsum(replace_na(Distance.x,0)),
                cumDist.y = cumsum(replace_na(Distance.y,0))) %>%
  ungroup()

cumdist_anitracker <- cumdist %>%
  group_by(Cow) %>% arrange(merge_index) %>%
  mutate(index = 1:n()) %>% ungroup() %>%
  ungroup() %>%
  select(Cow, index, cumDist.x, DistFlag.x) %>%
  rename(Flag = DistFlag.x,
                cumDist = cumDist.x) %>%
  mutate(Source = "animaltracker")

cumdist_manual <- cumdist %>%
  group_by(Cow) %>% arrange(merge_index) %>%
  mutate(index = 1:n()) %>% ungroup() %>%
  select( index, Cow, cumDist.y, DistFlag.y) %>%
  rename( Flag = DistFlag.y,
                cumDist = cumDist.y) %>%
  mutate(Source = "manual")

plot_data <- bind_rows(cumdist_anitracker, cumdist_manual)

ggplot(plot_data, aes(x=index, y=cumDist, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  ylab("Cumulative Distance") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```



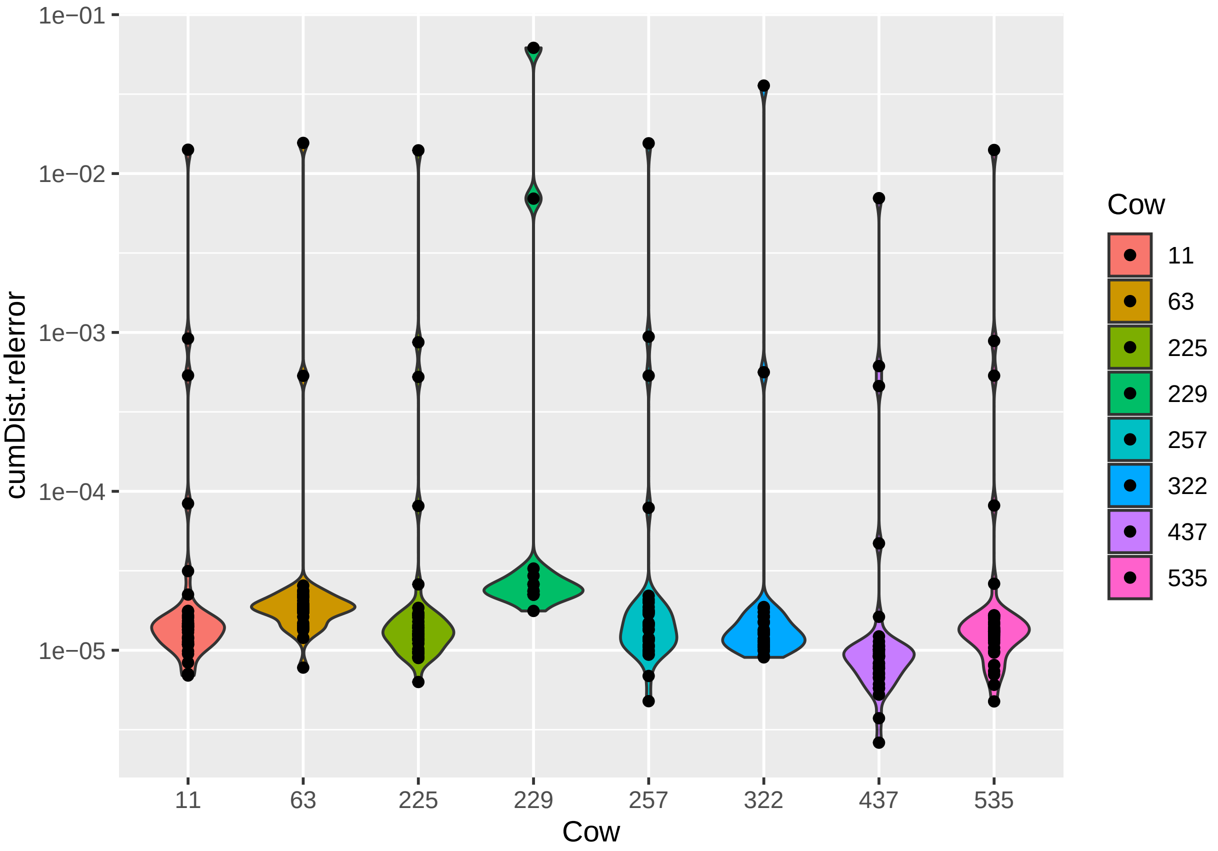
Relative Error of Cumulative Distance Estimates

The following summarizes the relative error of cumulative distances calculated from the `animaltracker` app in comparison to the manually processed data.

```
error_cumdist <- join %>%
  group_by(Cow) %>%
  arrange(merge_index) %>%
  mutate(
    Dist.y = lag(Distance.y,1),
    cumDist.x = cumsum(replace_na(Distance.x,0)),
    cumDist.y = cumsum(replace_na(Dist.y,0)) ) %>%
  group_by(Cow, Date.x) %>%
  summarize(
    cumDist.x = sum(cumDist.x, na.rm=TRUE),
    cumDist.y = sum(cumDist.y, na.rm = TRUE),
    cumDist.relerror = (cumDist.x-cumDist.y)/cumDist.y
  ) %>%
  ungroup()
```

Based on N = 243 days of data, the overall relative error rate in cumulative distance per day is 0.08%.

```
ggplot(error_cumdist, aes(x = Cow, y = cumDist.relerror, fill = Cow))+
    geom_violin(trim=TRUE)+
    geom_point()+
    scale_y_continuous(trans='log10')
```



```
error_cumdist %>%
  group_by(Cow) %>%
  mutate(index = 1:n()) %>%
  ungroup() %>%
  select(index, name = Cow, value = cumDist.relerror) %>%
  mutate(name = paste0("Cow_", name)) %>%
  pivot_wider() %>%
  select(-index) %>%
  psych::describe() %>%
  select(n, mean, sd, median, range, se ) %>%
  print(digits = 4)
```

```
##          n   mean     sd median  range     se
## Cow_11  35 0.0005 0.0024      0 0.0141 0.0004
## Cow_63  29 0.0006 0.0029      0 0.0156 0.0005
## Cow_225 35 0.0005 0.0024      0 0.0140 0.0004
## Cow_229 12 0.0058 0.0178      0 0.0618 0.0051
## Cow_257 35 0.0005 0.0026      0 0.0155 0.0004
## Cow_322 27 0.0014 0.0069      0 0.0358 0.0013
## Cow_437 35 0.0002 0.0012      0 0.0070 0.0002
## Cow_535 35 0.0005 0.0024      0 0.0141 0.0004
```

Rate of Travel

The following describes differences in estimated `Rate` measurements (meters/min) between the data cleaned in `animaltracker` and the manually cleaned data.

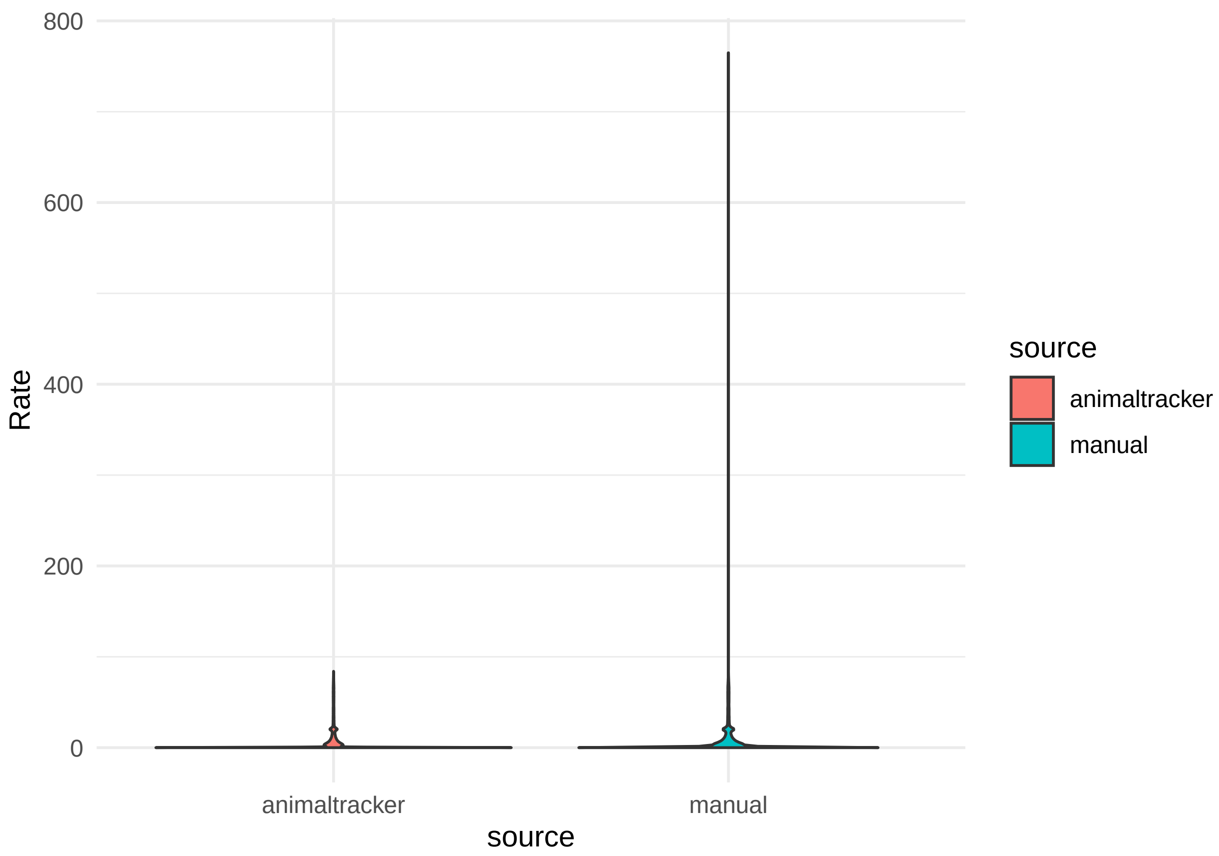
```
rates_keep <- join %>%
    filter(Keep.x > 0 ) %>%
    select(merge_index, Rate = Rate.x) %>%
    mutate(source = "animaltracker") %>%
  rbind(
    join %>%
      filter(Keep.y > 0) %>%
      select(merge_index, Rate = Rate.y) %>%
      mutate(source = "manual")
  ) %>%
  mutate(source = factor(source),
         Rate = as.numeric(Rate))

rates_keep %>%
  pivot_wider(names_from = "source", values_from="Rate") %>%
  select(-merge_index) %>%
  psych::describe() %>%
  select(n, mean, sd, median, range, se ) %>%
  print(digits = 3)
```

```
##                    n  mean     sd median   range    se
## animaltracker 164913 6.281 13.446      0  83.998 0.033
## manual        164970 6.297 13.696      0 764.775 0.034
```

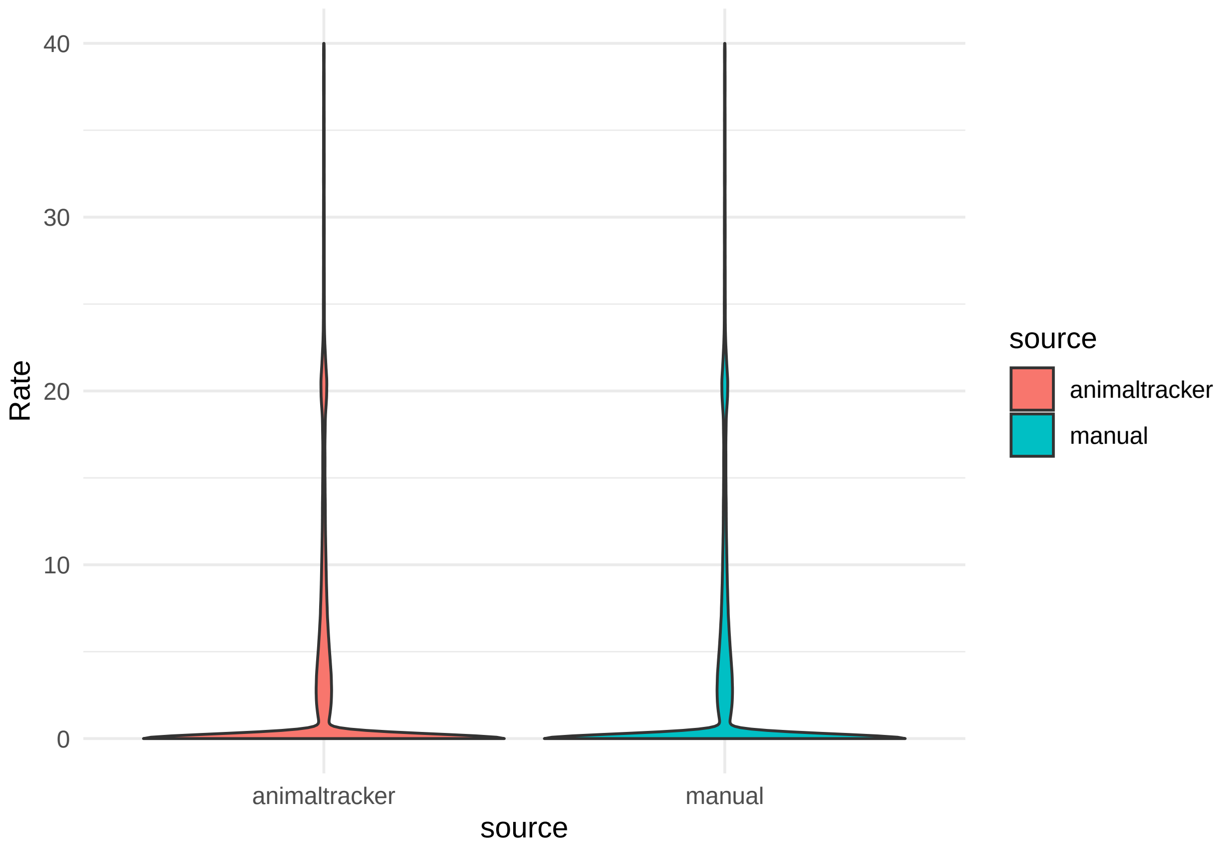
```
ggplot(rates_keep, aes(x = source, y = Rate, fill = source))+
  geom_violin(trim=TRUE) +
  theme_minimal()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_ydensity).
```



Restricting to the bulk of measured rates (< 40 meters/min), we see nearly identical distributions.

```r
ggplot(rates_keep %>% filter(Rate < 40), aes(x = source, y = Rate, fill = source))+
  geom_violin(trim=TRUE) +
  theme_minimal()
```


40
30
20
10
0
Rate
animaltracker
manual
source
source
animaltracker
manual